# Outils de traitement des langues et corpus spécialisés: l'exemple d'Unitex

Rosa Cetro

**Outils de traitement des langues et corpus spécialisés : l'exemple d'Unitex**

Rosa CETRO

Università di Brescia et Université Paris Est Marne-la-Vallée

**Résumé**

Il sera question dans cet article de dresser un panorama des outils de traitement de langues et de leur application à la terminologie, par le biais de corpus spécialisés. Dans la première partie, nous discuterons des apports de l'informatique à la pratique terminologique, en retraçant un bref aperçu historique des contacts entre ces deux disciplines jusqu'à nos jours. Les différents types d'approches utilisées (approches statistiques, linguistiques et hybrides) dans la conception des outils informatiques de traitement des langues feront l'objet de la deuxième partie de cet article. Dans la troisième partie, consacrée à Unitex, nous illustrerons des exemples d'utilisations de cet outil pour le travail terminologique. La comparaison d'Unitex avec d'autres outils constituera notre quatrième et dernière partie.

## 1. Terminologie et informatique : bref aperçu historique

La terminologie voit le jour dans les années 1930, grâce à l'œuvre de l'ingénieur autrichien Eugen Wüster, qui formule la *Théorie Générale de la Terminologie*, et présente la terminologie comme une matière interdisciplinaire, au carrefour entre linguistique, sciences cognitives, sciences de l'information, sciences de la communication et informatique. Dans les trente premières années de cette nouvelle discipline, c'est surtout son caractère systématique qui s'affirme. Le rôle joué par l'informatique est, à cette époque, très limité. Disons, avec P. Drouin **(DROUIN 2002 : 43)** que « *entre les années 1930 et 1960 la terminologie fournit […] à l'informatique une réflexion théorique sur les concepts et leur gestion* ». Dans la décennie 1960, et pour une bonne partie de la décennie 1970, la contribution de l'informatique consiste surtout dans la mise en place des premières grandes banques de terminologie et des fichiers de taille imposante, réalisés pour des projets gouvernementaux d'aménagement linguistique (comme la banque de données DICAUTOM, en 1963) ou pour gérer l'information d'importantes entreprises privées. Lorsque Wüster écrivait **(WÜSTER 1974 : 98)** :

> « Die Informatik ist von allen bisher besprochenen Wissenschaften die jüngste. Es ist die Wissenschaft vom Bau und von der Verwendung der Elektronenrechner. Die Elektronenrechner dienen nicht nur für mathematische Rechnungen. Sie sind auch das wichtigste technische Hilfsmittel einer anderen, nicht sehr viel älteren Wissenschaft, nämlich der Dokumentations- und Informationswissenschaft. Der Name "Informatik" soll das wohl andeuten »[1],

il ne pouvait pas imaginer à quel point les rapports entre terminologie et informatique devaient changer. Les banques de terminologie passent sur cédérom dans les années 1980, avec la naissance de la *terminotique* (terme créé à partir de *terminologie* et de *bureautique*). L'apparition des premiers outils pour le traitement des textes remonte au début des années 1990. Il s'agit de logiciels conçus principalement pour trois tâches : l'extraction

[1] « De toutes les sciences dont nous avons parlé jusqu'ici, l'informatique est la plus jeune. C'est la science de la construction et de l'utilisation des ordinateurs. Les ordinateurs ne servent pas uniquement à effectuer des opérations mathématiques. Ils constituent également une aide technique capitale pour une autre science qui n'est pas beaucoup plus vieille que l'informatique, la science de la documentation et de l'information, d'où le nom d'informatique », dans : RONDEAU Guy, FELBER Helmut (1981 : 102).

terminologique, la structuration d'ontologies et l'alignement terminologique (dans le cas de travaux bilingues), dont nous parlerons plus en détail à la section 2.

Si en 1960 le rôle de l'informatique en terminologie est limité au stockage de l'information, comme en témoigne la création des banques de terminologie, dans la décennie 1990, elle devient désormais indispensable au terminologue, qui se voit déchargé d'une partie des tâches répétitives de son travail. La rapidité dans le dépouillement des textes, la création de nouvelles ressources terminologiques et la facilité de mise à jour de ressources existantes sont à la base de l'essor de l'informatique en terminologie.

## 2. Approches utilisées dans la conception des outils informatiques

Avant de passer à l'analyse des approches utilisées dans la conception de ces outils, il nous semble nécessaire d'évoquer le contexte dans lequel ces recherches ont vu le jour. L'incitation fondamentale à mener ces travaux est plus venue, dans un premier temps, du monde de l'entreprise et de ses besoins connexes, que du monde académique. Cela pour plusieurs raisons : tout d'abord, la disponibilité croissante de corpus textuels en format électronique ; de plus, la nécessité pour les entreprises de gérer une grosse quantité d'informations et d'avoir à disposition des produits terminologiques, tels que des glossaires, des thésaurus ou des ontologies. Si l'on considère aussi que les entreprises ont tout intérêt à garder la confidentialité de certaines des données contenues dans ces documents, il ne faut pas s'étonner que bon nombre des projets de logiciels conçus pour des applications terminologiques aient été menés dans le cadre de projets privés financés par les entreprises. Toutefois, d'importants projets ont aussi été menés au sein du monde académique, comme le progiciel Termino (voir section 2.2). Une motivation supplémentaire pour ces recherches est venue aussi du renouveau des travaux en analyse statistique de la langue.

Un autre aspect à ne pas négliger est l'influence que le monde francophone a eue sur ces recherches, comme le soulignent Didier Bourigault et Christian Jacquemin **(BOURIGAULT et JACQUEMIN 2000 : 217-218)** :

> « On peut voir deux explications au fait que le français, et non l'anglais, soit privilégié, au moins à l'origine, dans les travaux en acquisition de terminologie à partir de corpus. D'abord la tâche de repérage automatique de syntagmes nominaux terminologiques est réputée être plus difficile pour le français, et les langues romanes, que pour l'anglais, ou l'allemand, de l'avis en particulier des chercheurs travaillant sur l'acquisition terminologique bilingue. Contrairement à l'anglais, qui construit ses termes complexes essentiellement par juxtaposition d'unités lexicales pleines, le français use abondamment des prépositions et des déterminants, ce qui rendrait la distinction entre terme et syntagme libre moins marquée a priori et plus difficile à saisir par des outils automatiques. Ensuite, c'est d'abord dans le contexte de la traduction et celui, lié, de l'aménagement linguistique que se fait sentir de façon la plus cruciale le rôle éminemment stratégique de la terminologie. Or les pays concernés en premier chef par ces problèmes ne sont pas par définition ceux où l'anglais est la seule langue officielle. »

Les progrès de l'informatique vont de pair avec une réflexion théorique sur la nature du terme, qui se traduit par un changement de perspective et s'éloigne de la rigidité du schéma wüstérien sur plusieurs aspects. Déjà, il ne s'agit plus, comme le préconisait Wüster, de découvrir *la* terminologie d'un domaine, donnée par avance et dont les experts du domaine sont les dépositaires, mais de procéder, par le biais de la documentation spécialisée, à la construction d'*une* terminologie possible de ce domaine. De même, c'est l'idée de la monoréférentialité du terme qui est mise en discussion, remplacée par la notion de variation terminologique : une unité terminologique peut présenter une ou plusieurs variantes

(syntaxiques ou morphosyntaxiques) et correspondre néanmoins à un seul et même concept. Sur le plan pratique, la prise en considération de l'existence des variantes terminologiques a entraîné une plus grande attention envers la syntaxe pour élaborer beaucoup de logiciels destinés à des applications terminologiques.

Parmi les tâches terminologiques visées par ces outils (citées à la section 1), la principale s'avère être l'acquisition automatique des termes à partir de corpus textuels. Les approches utilisées dans la conception de ces logiciels se répartissent essentiellement entre approches statistiques et approches linguistiques. Il existe aussi des approches hybrides, nées de la combinaison de méthodes statistiques et de méthodes linguistiques.

### 2.1. Les approches statistiques

Les approches statistiques trouvent leur origine dans les modèles mécaniques utilisés en documentation à la fin des années 1980. Basés sur des algorithmes, ces modèles prévoyaient l'exploitation de la force brute des ordinateurs pour identifier les collocations dans les gros corpus, sur le critère de la répétition de segments de texte. En France, c'est le cas, par exemple, des travaux de statistique textuelle de Lebart et Salem (1988) **(DROUIN 2002 : 58-60)**.

Les avantages reconnus aux approches statistiques sont essentiellement la capacité de traiter des corpus de taille imposante et l'indépendance de ressources linguistiques (telles que grammaires ou dictionnaires) extérieures au corpus traité. Ce dernier aspect rend les techniques statistiques plus rapides et aussi plus économiques, car les ressources linguistiques, comme fruit d'un travail manuel, sont souvent coûteuses.

Dès lors que l'extraction terminologique se fonde sur le critère de la fréquence, un risque des approches statistiques, au détriment de la qualité des résultats, est de passer sous silence les termes dont les occurrences sont peu nombreuses. De plus, les performances de ce type d'outils sont bien plus intéressantes sur des corpus de grande taille, si l'on en croit Patrick Drouin **(DROUIN 2002 : 93)** : « *On considère généralement que l'application de techniques statistiques à des corpus de taille inférieure à 100 000 occurrences ne conduit pas à l'obtention de résultats fiables et justifiables* ».

Un outil conçu sur un modèle statistique est le logiciel ANA (Apprentissage Naturel Automatique), développé en 1992 au sein du Centre d'études atomiques (CEA) de Cadarache par Chantal Enguehard. Basé sur deux méthodes algorithmiques, ANA est un outil multilingue[2]. Le système accepte de traiter des données brutes, non étiquetées préalablement. L'extraction terminologique prévoit deux phases. Dans un premier module appelé « familiarisation », le logiciel procède à l'extraction des connaissances dans le corpus sous forme de quatre listes, en séparant les mots fonctionnels (conjonctions, adverbes, etc.) des candidats termes. Cette liste de candidats termes (*amorçage*) est ensuite enrichie dans le second module, « découverte », sur la base des cooccurrences repérées dans le corpus **(ENGUEHARD 1993 : 376-377)**.

### 2.2. Les approches linguistiques

Dans cette catégorie entrent deux types de systèmes : les systèmes exploitant des informations syntaxiques et les systèmes qui utilisent des informations lexicales ou morphologiques. D'après la distinction de P. Drouin **(DROUIN 2002 : 66)** :

[2] ANA est limité aux langues non agglutinantes et jusqu'à ce jour a été testé sur le français, l'anglais et l'italien.

« Les premiers reposent sur une analyse complète de la phrase en ses constituants afin d'en dégager les syntagmes intéressants selon les objectifs de la recherche. Dans le second cas, des grammaires locales procèdent à une analyse de surface de la phrase à la recherche de syntagmes potentiels ».

Les avantages des techniques linguistiques sont une description linguistique fine et la possibilité de traiter des corpus de petite taille.

Termino, développé par S. David et P. Plante en 1990, dans le cadre d'une collaboration entre l'Office de la Langue Française du Québec et l'Université du Québec à Montréal, a été le tout premier outil pour l'extraction terminologique. Son but était de repérer dans un corpus les unités nominales syntaxiques susceptibles de se lexicaliser (suivant la théorie des *synapsies* de Benveniste). Dix ans plus tard, des modifications sur Termino ont donné naissance à Nomino, système de dépouillement terminologique.

Parmi les outils exploitant des critères syntaxiques, nous citons Lexter et Fastr. Le premier, développé par D. Bourigault en 1993 pour EDF et désormais propriété de cette dernière, associe deux tâches : l'extraction terminologique et la structuration de terminologie. Il extrait les candidats termes à partir d'un corpus préalablement étiqueté et désambiguïsé, et il les organise dans un deuxième temps dans un réseau sémantique. Le deuxième, développé par Ch. Jacquemin en 1996, est un analyseur syntaxique, dont l'enjeu principal est la reconnaissance des variantes terminologiques (qui peuvent être syntaxiques, morphosyntaxiques et sémantico-syntaxiques). Ce logiciel peut être utilisé en association avec Lexter : il accepte en entrée une liste de candidats termes extraits par ce dernier, et il en fournit en sortie les variantes terminologiques **(BOURIGAULT et JACQUEMIN 2000 : 225-226)**.

Dans la catégorie des outils fondés sur des approches linguistiques entre aussi le logiciel Unitex, dont nous parlerons en détail à la section 3.

### 2.3. Les approches hybrides

Il existe aussi une troisième catégorie d'outils qui associent des techniques statistiques à des techniques linguistiques et qui sont en recrudescence dans la dernière décennie. En général, les techniques linguistiques utilisées dans les approches hybrides ont recours à l'analyse syntaxique.

L'ordre d'application des différents types de techniques varie selon les outils. Dans le cas d'ACABIT, par exemple, conçu par B. Daille pour la société IBM en 1994, un corpus textuel - préalablement étiqueté - est d'abord soumis à un traitement linguistique à l'aide de transducteurs, qui en dégagent des séquences nominales et les ramènent à des candidats termes binaires. Les résultats sont soumis dans un deuxième temps à un filtrage statistique **(BOURIGAULT et JACQUEMIN 2000 : 224-225)**. En revanche, le logiciel Xtract, développé en 1993 par F. Smadja, prévoit comme première étape le filtrage statistique et comme deuxième étape l'application d'analyses linguistiques **(BOURIGAULT et JACQUEMIN 2000 : 225)**.

A l'instar des approches statistiques, dont elles partagent la systématicité, la rapidité et l'indépendance par rapport aux domaines abordés dans les corpus, les approches hybrides s'avèrent plus performantes sur les gros corpus que sur les corpus de petite taille.

## 3. Unitex : exemples d'utilisations pour le travail terminologique

Développé par S. Paumier en 2002, Unitex n'est pas un outil expressément conçu pour la terminologie, mais un outil de traitement de textes, sur le modèle du logiciel Intex,

développé par M. Silberztein au LADL (Laboratoire d'Automatique Documentaire et Linguistique) en 1994. Toutefois, certaines de ses fonctions peuvent aussi être exploitées pour le travail terminologique, comme nous le verrons à partir d'une série d'expériences de traitement textuel sur un corpus spécialisé ayant trait à la médecine thermale, qui constitue le corpus de référence de notre thèse.

Comme nous l'avons déjà annoncé au paragraphe 2.2, Unitex est un outil basé sur des techniques linguistiques, en particulier des ressources lexicales et morphologiques : il s'agit des dictionnaires électroniques et des tables de lexique-grammaire élaborés au LADL de l'Université Paris VII.

Lorsqu'on soumet un texte au logiciel, il exécute plusieurs opérations, affichées dans des fenêtres différentes : l'extraction des formes (tokens) repérées dans le texte, le découpage en phrases, l'étiquetage morphosyntaxique des formes[3].

### 3.1. Le comptage des formes

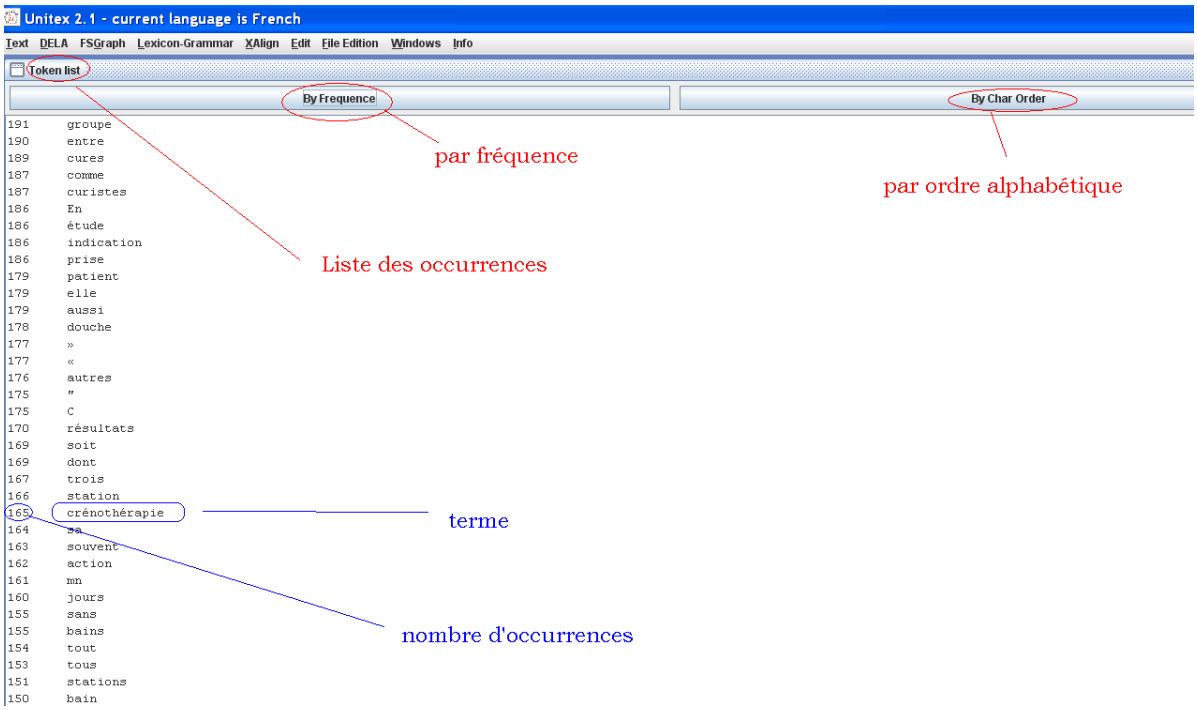


Figure 1 : fenêtre Token list.

La figure 1 montre l'extraction des formes d'un texte (Token list) : on peut choisir d'afficher cette liste par fréquence (by frequence) ou par ordre alphabétique (by char order). Chaque forme repérée dans le texte est affichée sur une ligne, accompagnée de son nombre d'occurrences.

### 3.2. Ressources lexicales : les dictionnaires

Dans une autre fenêtre (Word Lists, figure 2), il est possible de voir les résultats de l'application des ressources lexicales et morphologiques du logiciel : les dictionnaires électroniques. Les formes contenues dans la Token list sans informations supplémentaires

[3] L'étiquetage morphosyntaxique fait par Unitex est sans levée des ambiguïtés.

sont ici accompagnées d'étiquettes morphologiques classées dans trois encadrés différents. Dans l'encadré en haut à gauche, sont listées les formes simples reconnues par le DELAF (Dictionnaire des formes simples), tandis que l'encadré en bas à gauche montre les formes polylexicales reconnues par le DELACF (Dictionnaire des formes composées). L'encadré de droite, en revanche, liste les formes du texte qui ne sont pas reconnues par les dictionnaires électroniques, ce qui peut constituer un point de départ pour l'identification de néologismes ou de termes spécifiques à un domaine. En ce qui concerne les entrées des dictionnaires, des couleurs différentes sont utilisées pour le codage des informations. L'entrée est en bleu, suivie d'une virgule et éventuellement du lemme auquel elle est rattachée, qui en revanche est en rouge. Les codes traduisant la catégorie grammaticale et le niveau de langue sont en vert, suivi des deux points « : » ; derrière ces derniers se situent les informations sur la flexion (genre et nombre, temps/mode, personne), en orange. Les tokens non reconnus par les dictionnaires sont listés en noir.

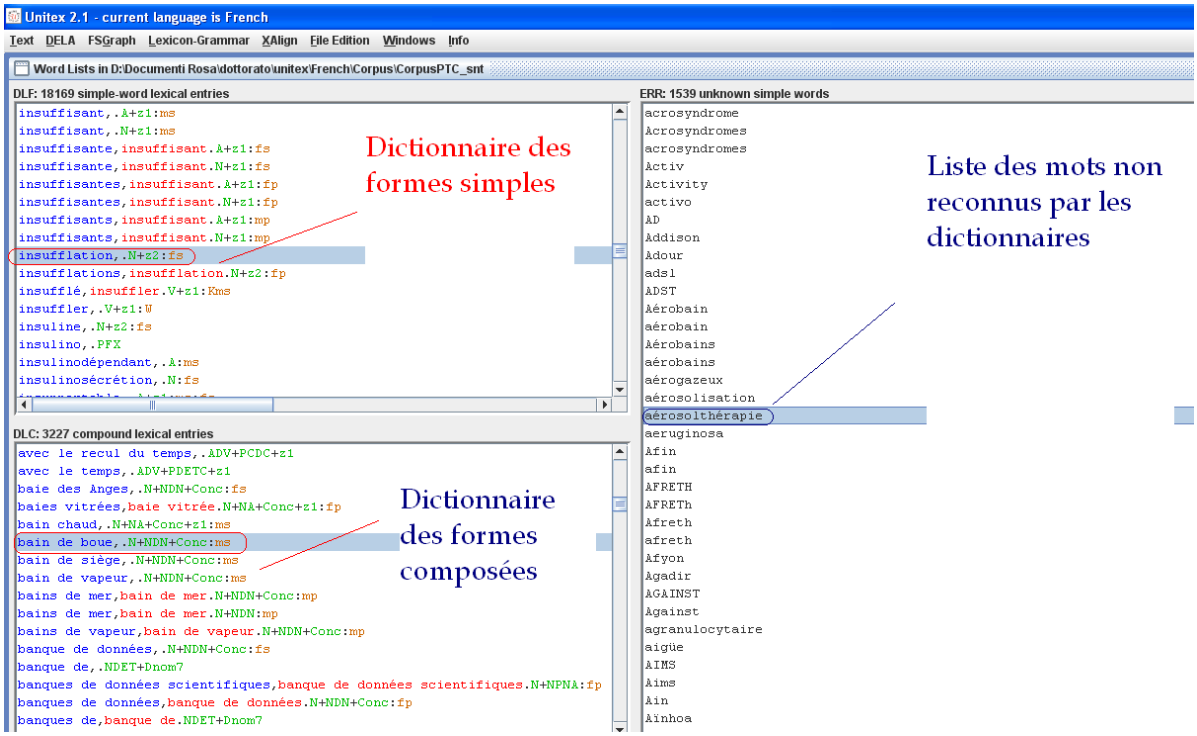


Figure 2 : fenêtre Word Lists.

### 3.3. Construction de ressources lexicales

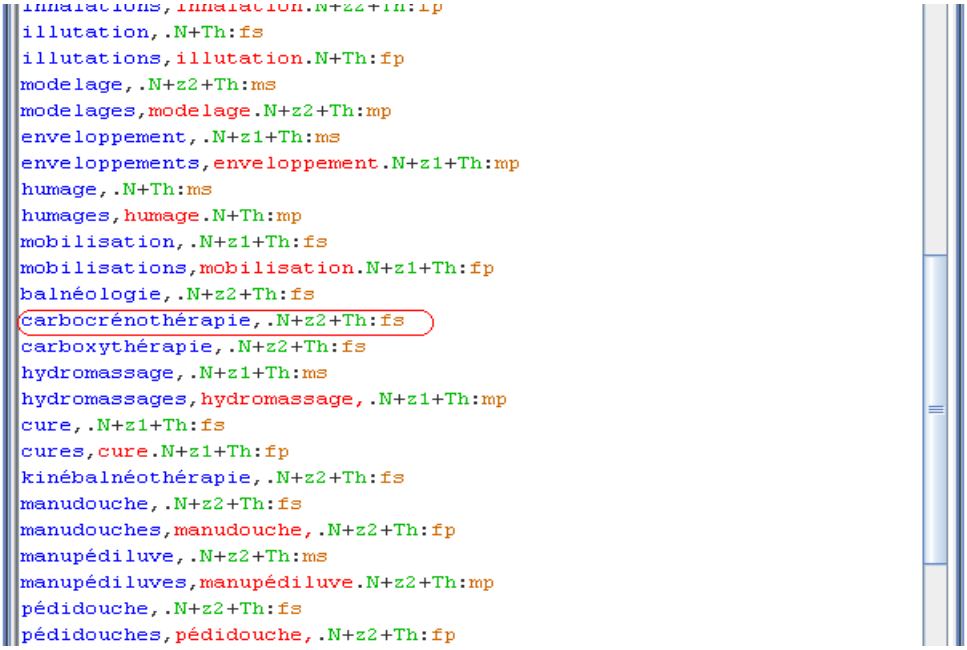


Figure 3 : le dictionnaire électronique balnéo simple.

Il est possible d'enrichir les ressources lexicales d'Unitex, par exemple en créant un dictionnaire électronique. C'est ce que nous avons fait avec le dictionnaire balnéo_simple.bin, dans lequel nous avons listé les termes simples désignant des soins en médecine thermale (figure 3). Chaque terme est répertorié selon la procédure que nous avons vue à la section 3.2, lorsque nous avons parlé du DELAF et du DELACF. Prenons, par exemple, le terme *carbocrénothérapie*, absent de la nomenclature du DELAF et codé dans notre dictionnaire de la façon suivante : carbocrénothérapie,.N+z2+Th:fs. Ces codes traduisent les informations suivantes : *cartbocrénothérapie* est un nom (N) qui n'est pas aussi fréquent qu'un mot de la langue générale (z2) et ses marques de genre et nombre sont « féminin » et « singulier » (fs). Nous avons choisi d'utiliser le code Th (Thermalisme) pour différencier les formes de notre dictionnaire de celles contenues pour le DELAF : cela s'explique par la volonté de marquer les formes nominales à partir desquelles sont créés les noms composés de la médecine thermale et par la possibilité de les utiliser dans des grammaires locales, comme nous le verrons à la section suivante (3.4.).

### 3.4. Recherche de contextes : expressions régulières et grammaires locales

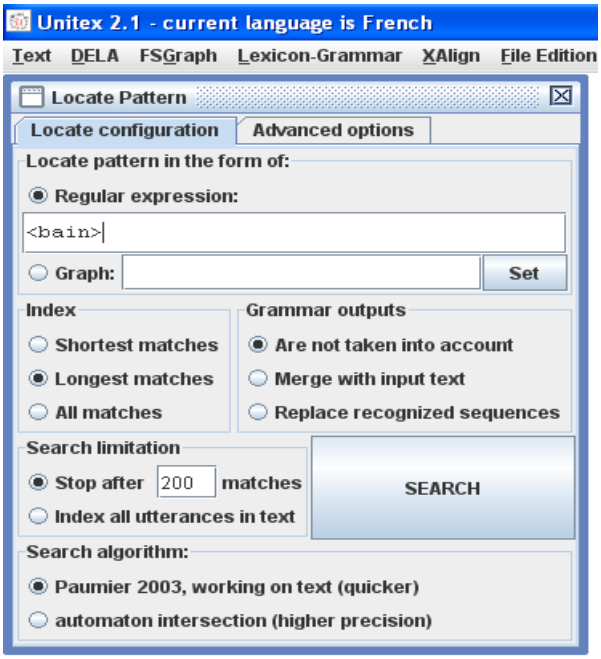


Figure 4 : la fonction Locate Pattern.

En terminologie, la recherche de contextes d'utilisation d'un terme est une tâche très importante, surtout pour réaliser des produits terminologiques comme des banques de données, des dictionnaires spécialisés ou des thésaurus. L'informatique a beaucoup facilité cette tâche aux terminologues, en rendant le dépouillement des sources bien plus rapide. Dans Unitex, la recherche de contextes d'utilisation se fait par le biais de la fonction Locate Pattern (figure 4), dans le menu Text. Il est possible de mener cette recherche de deux façons : par expression régulière (regular expression) ou par l'application de grammaires locales (graph). Pour la recherche de contextes à partir d'une expression régulière, il suffit de cocher la case « Regular expression » et d'écrire le mot que l'on veut rechercher. Nous voulons trouver, par exemple, toutes les occurrences du mot *bain*. L'utilisation des chevrons (< >) permet le repérage non seulement de *bain*, mais aussi de *bains* ; c'est-à-dire qu'une requête avec un lemme entre chevrons provoque la recherche de toutes les formes fléchies de ce lemme dans un texte. Seules les formes lemmatisées peuvent donc être insérées entre chevrons : il n'en va pas de même pour les formes fléchies et pour les formes non reconnues par les dictionnaires du programme, pour lesquelles la recherche doit impérativement être lancée sans l'utilisation des chevrons. En cochant la case « Graph », la recherche de contextes se fait à partir d'une grammaire locale, élaborée par un utilisateur du logiciel.

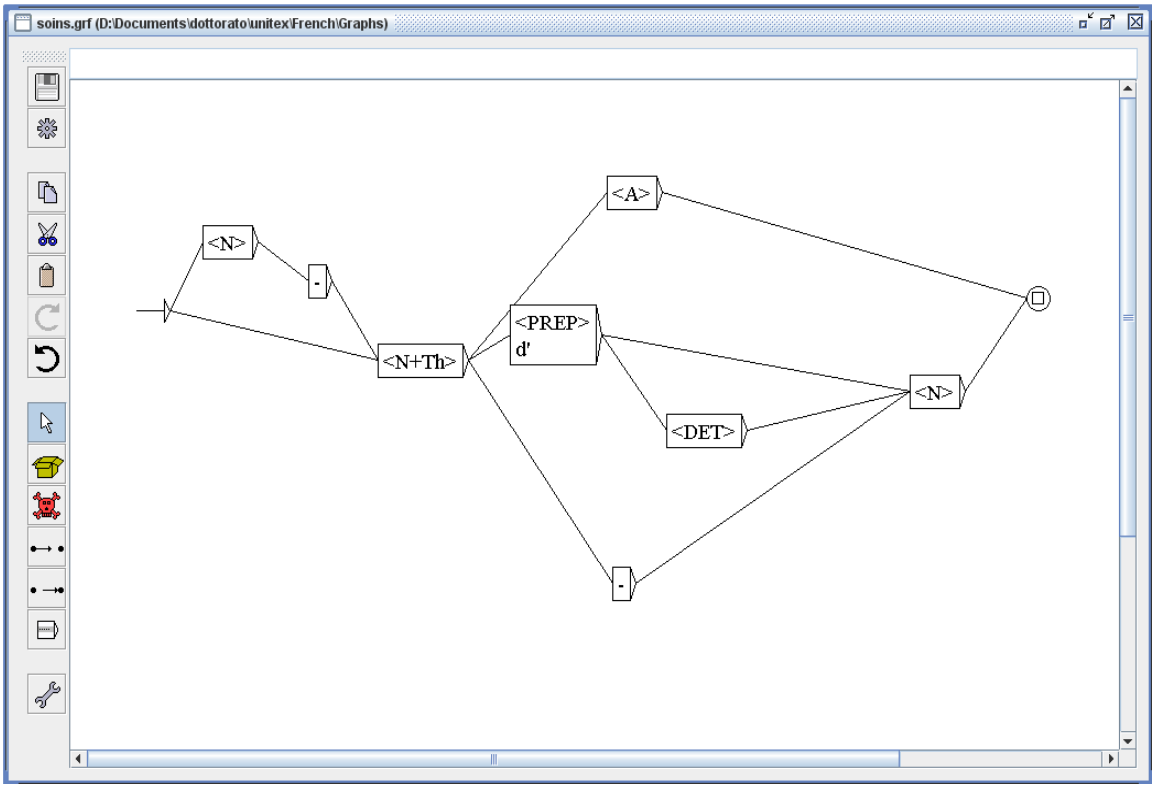


Figure 5 : le graphe Soins, exemple de grammaire locale.

Les grammaires locales, qui prennent la forme de graphes, sont des représentations par automates de structures linguistiques. Elles sont utilisées pour repérer dans un corpus textuel des segments de texte, à partir d'informations syntaxiques et/ou lexicales. La figure 5 montre le graphe « Soins », qui est un exemple de grammaire locale à mi-chemin entre informations lexicales et syntaxiques. Le but de ce graphe est d'identifier des noms composés désignant les soins utilisés en médecine thermale à partir d'un corpus spécialisé. La case <N+Th> permet de reconnaître toutes les formes simples enregistrées dans notre dictionnaire balnéo_simple, que nous avons vu à la section 3.3. Ce graphe reconnaît :

- Les séquences de deux noms, séparés ou non par un tiret, dont l'un au moins est un nom technique : ex. *humage-nébulisation* ;
- Les séquences d'un nom technique et d'un adjectif (si l'adjectif est dans les dictionnaires utilisés), comme *douche filiforme* et *insufflations tubaires* ;
- Les séquences composées d'un nom technique, une préposition et un autre nom (qu'il soit précédé ou non d'un déterminant), comme *cure de boisson* ou *injection de gaz*.

Dans les deux cas (recherche par expression régulière ou par grammaire locale), les concordances obtenues sont affichées en bleu, sous formes de liens hypertextuels (figure 6). Il suffit ensuite de cliquer sur ces liens pour retrouver le contexte d'utilisation de la forme recherchée, qui apparaît surlignée en bleu (figure 7).

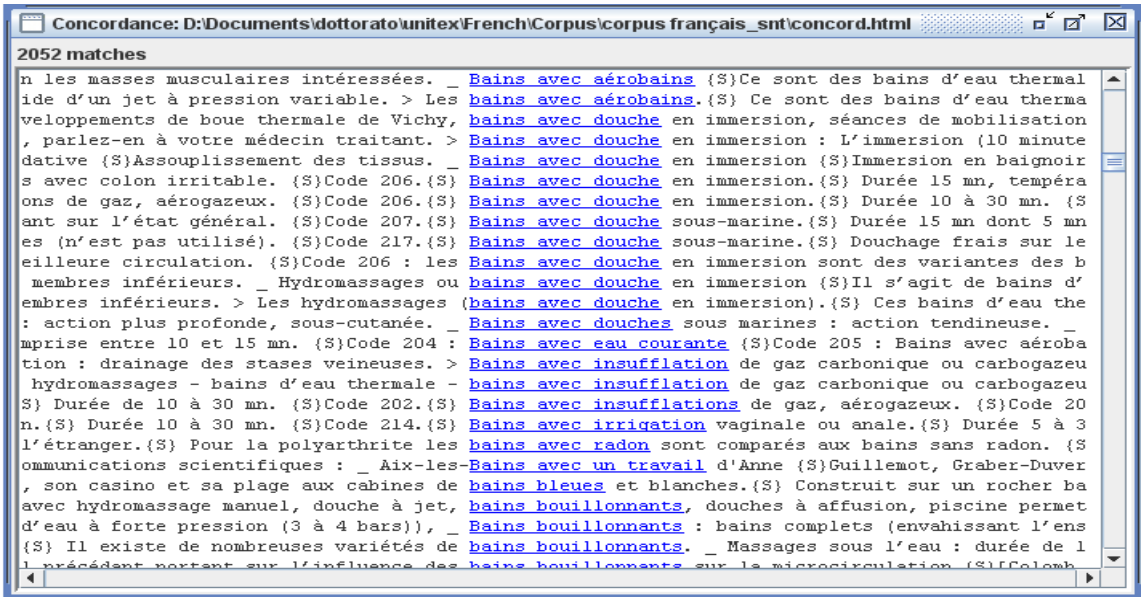


Figure 6 : affichage des concordances.

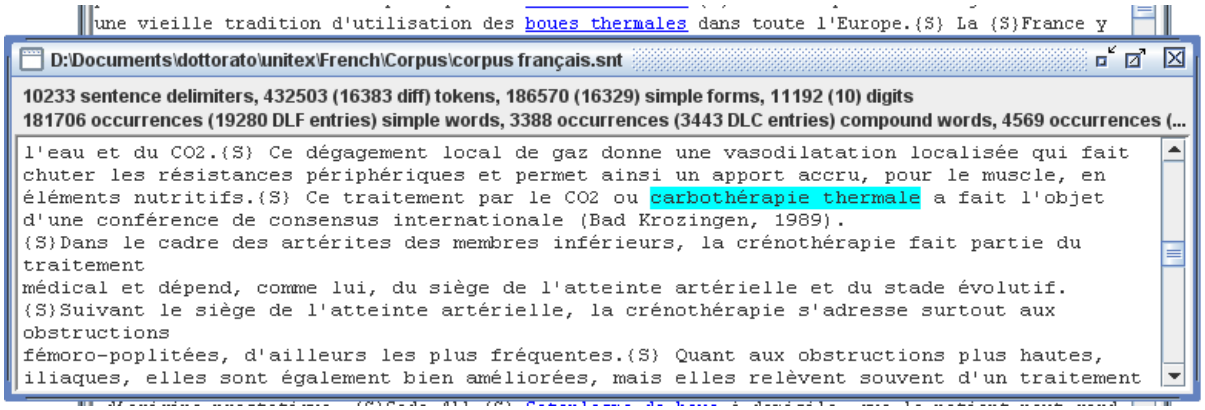


Figure 7 : affichage du contexte repéré.

### 3.5. X-Align : le programme d'alignement bilingue

X-Align est un programme d'alignement bilingue mobilisable depuis Unitex, qui, en tant qu'un outil multilingue est disponible pour l'anglais, le français, l'italien, le finnois, l'espagnol, l'allemand, le géorgien ancien, le grec (ancien et moderne), l'arabe, le norvégien, le portugais (variétés du Brésil et du Portugal), le polonais, le russe, le serbe et le thaï. La fonction de X-Align est de mettre en correspondance un texte en deux langues différentes. Chaque texte est découpé en segments, qui peuvent être des phrases ; et chaque segment est associé à un ou plusieurs autres segments. Si nous disposions de corpus parallèles de la médecine thermale pour notre étude, nous pourrions l'utiliser pour l'alignement terminologique. Un corpus parallèle est un corpus de mêmes textes en deux langues différentes. Toutefois, comme nous disposons de corpus comparables mais non parallèles, c'est-à-dire de corpus de textes traitant du même sujet en deux langues différentes, nous ne pouvons pas exploiter cette fonction.

## 4. État de l'art : comparaison avec d'autres outils

Comme une partie de notre travail de thèse porte sur l'évaluation du logiciel Unitex, nous avons testé d'autres outils pour avoir des termes de comparaison. Dans ce qui suit, nous relatons les difficultés rencontrées et les résultats obtenus jusqu'à présent.

### 4.1. Difficultés rencontrées

Plusieurs facteurs nous ont empêchée de tester bon nombre des outils décrits à la section 2 : nous les traitons en détail dans ce paragraphe.

La rapidité avec laquelle ces outils évoluent explique l'indisponibilité de certains d'entre eux. Tel a été le cas de Terminology Extractor, outil pour l'extraction terminologique travaillant sur l'anglais et le français, développé par Etienne Cornu pour l'entreprise Chamblon Systems Inc. Cambridge (Ontario, Canada). Comme le téléchargement n'a pas donné de résultats, nous avons contacté l'auteur, qui nous a informée de l'indisponibilité de Terminology Extractor. La même situation s'est avérée pour l'outil Mantex, conçu pour le système d'exploitation Macintosh en 2000 par P. Frath. Cet outil, fondé sur des techniques statistiques et visant l'identification des syntagmes répétés dans un corpus, est désormais obsolète[4].

Comme nous l'avons vu à la section 2, plusieurs outils de traitement de texte ont été développés dans le cadre de projets destinés aux entreprises, comme Lexter de D. Bourigault (1993). Ce logiciel, dont la propriété est désormais détenue par Électricité De France (EDF), n'a pas pu être testé pour cette raison.

Dans d'autres cas, les obstacles ont été l'insuffisance des manuels d'installation et d'utilisation de ces logiciels, limités à un fichier « Read me » de deux pages et évidemment adressés à des professionnels chevronnés : nous nous référons aux outils Fastr de Ch. Jacquemin et ACABIT de B. Daille. En ce qui concerne ACABIT, nous avons rencontré un empêchement supplémentaire : le système n'accepte que des données prétraitées. Ce qui implique le recours à des programmes extérieurs au logiciel et par conséquent des temps plus longs pour l'obtention des résultats.

### 4.2. Résultats obtenus

À ce jour, nous avons pu tester deux outils : l'un est ANA (Apprentissage Naturel Automatique), l'autre Cordial Analyseur. Le premier, dont nous avons parlé à la section 2.1., est un logiciel d'extraction terminologique fondé sur des approches statistiques, qui n'a jamais été distribué. Néanmoins, il est possible de le tester en contactant l'équipe de C. Enguehard, qui continue à travailler à ce projet à l'Université de Nantes. Le deuxième, tout comme Unitex, n'est pas un outil conçu expressément pour la terminologie. Il a été développé par P. Séguéla, au sein de la société Synapse. Cordial Analyseur est un analyseur syntaxique à base d'approches hybrides, qui mêle un filtrage statistique à l'application de ressources verbales construites à partir des tables de verbes de lexique-grammaire de Maurice Gross, et d'un dictionnaire de collocations. Bien que ce logiciel soit payant, il est possible de le tester gratuitement[5] par mail.

[4] http://www.atala.org/-Outils-pour-le-TAL-

[5] http://www.synapse-fr.com/Cordial_Analyseur/Presentation_Cordial_Analyseur.htm.

Pour le test des deux outils, nous avons soumis à l'analyse : le texte *B_019 Alger*, de petite taille (environ 1.500 mots), et le fichier *Corpus français*, qui rassemble tous les textes du corpus d'analyse de notre thèse et qui compte environ 200 000 mots, dont *B_019 Alger* est un échantillon.

Le texte *B_019 Alger* n'a pas pu être traité par ANA en raison de sa petite taille, ce qui confirme un point faible des approches statistiques. Les résultats du traitement du fichier *Corpus français* nous ont été fournis sous forme d'un tableau de texte. Ils sont organisés dans trois colonnes : dans la première, les candidats termes extraits ; dans la deuxième, le nombre d'occurrences ; dans la troisième, les segments de texte d'où le candidat terme a été extrait, avec le nombre d'occurrences pour chaque segment :

| | | |
|---|---|---|
| Cure de boisson | 31 | (cure de boisson, 30)<br>(cures de boisson, 1) |

L'évaluation de ces résultats est en cours de réalisation.

En ce qui concerne Cordial Analyseur, tant *B_019 Alger* que *Corpus français* ont pu être traités, et les résultats nous ont été fournis dans un fichier ouvrable avec tout programme d'édition de textes. Les textes ont été segmentés par phrases et à chaque forme une étiquette indiquant la catégorie grammaticale a été assignée. Après ce découpage en phrases, on nous fournit les statistiques de l'analyse du texte, qui reconnaissent les pourcentages de fréquence des parties du discours dans le texte (verbes, adjectifs, substantifs, etc.). Bien que, à l'instar d'Unitex, Cordial Analyseur permette d'extraire toutes les formes contenues dans un texte, certains aspects semblent suggérer qu'il n'est pas particulièrement adapté au travail terminologique. Les unités terminologiques ont souvent la forme d'unités polylexicales, dont la fixité est variable : si la recherche des contextes est limitée aux collocations contenues dans un dictionnaire, comme dans le cas de Cordial Analyseur, on ne peut pas repérer les collocations qui ne sont pas répertoriées dans le dictionnaire du logiciel. De plus, il n'est pas possible d'intégrer ses propres ressources lexicales dans Cordial Analyseur, à la différence de ce que nous avons pu voir avec Unitex. Le logiciel est limité au français, donc on ne peut pas l'exploiter pour des travaux de terminologie bilingues ou plurilingues.

### 4.3. En guise de conclusion : Unitex pour le travail terminologique ?

Sur la base de ce que nous avons analysé aux sections 3 et 4, nous pouvons apporter une réponse, au moins provisoire, à la question suivante : peut-on considérer Unitex comme un outil de support au travail terminologique ? Nous avons vu que, bien qu'Unitex n'ait pas été conçu expressément pour la terminologie, il permet l'accomplissement de certaines tâches de type terminologique : l'extraction terminologique, la détection de néologismes et la création de ressources lexicales, la recherche de contextes sur la base de critères variés, l'application de grammaires locales pour la recherche de motifs, la fonction d'alignement bilingue sur corpus parallèles. Si l'on ajoute à cela les avantages sur les approches statistiques, c'est-à-dire la possibilité de traiter des petits corpus textuels et le fait que toutes les formes d'un corpus sont listées (même celles dont la fréquence est 1), notre réponse à la question ci-dessus sera plutôt positive.

### Références bibliographiques

BOURIGAULT Didier et JACQUEMIN Christian (2000), « Construction de ressources terminologiques », in PIERREL J.-M. (dir.), *Ingénierie des langues*, Paris, Hermès, p. 215-233.

CABRÉ Maria Teresa (1998), *La terminologie. Théorie, méthodes et applications*, Paris, A. Colin.
CORI Marcel, DAVID Sophie, LÉON Jacqueline (2008), « La construction des faits en linguistique : la place des corpus », *Langages*, 171.
DROUIN Patrick (2002), *Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés*, thèse de doctorat, Université de Montréal.
ENGUEHARD Chantal (1993), « Acquisition de terminologie à partir de gros corpus », *Informatique et Langue Naturelle*, ILN'93, Nantes, p. 373-384.
L'HOMME Marie-Claude (2002), « Nouvelles technologies et recherche terminologique. Techniques d'extraction des données terminologiques et leur impact sur le travail du terminographe », lien : http://olst.ling.umontreal.ca/pdf/textHomme.pdf.
L'HOMME Marie-Claude (2004), *La terminologie : principes et pratiques*, Montréal, Les Presses de l'Université de Montréal.
LAPORTE Éric (2000), « Mots et niveau lexical », in PIERREL J.-M. (dir.), *Ingénierie des langues*, Paris, Hermès, p. 1-29.
PAUMIER Sébastien (2009), *Unitex 2.1. User manual*, téléchargeable à l'adresse suivante : http://www-igm.univ-mlv.fr/~unitex/index.php?page=4.
RONDEAU Guy, FELBER Helmut (dir.) (1981), *Textes choisis de terminologie*, Groupe interdisciplinaire de recherche scientifique et appliquée en terminologie, Université Laval.
WÜSTER Eugen (1974), « Die Allgemeine Terminologielehre – ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften », *Linguistics*, 119, p. 61-106.